SEIS | 734

Project Plan for the Million Song Recommendation Engine

Jason Baker, Joshua Janzen, Nomvelo Moyo & Usman Waheed

Graduate Programs in Software
Spring 2015

University of St. Thomas, St. Paul, Minnesota

Problem Statement

The primary objective of our project is to devise a model for a music recommendation engine, specifically in the context of popular songs and listening patterns stored in the Million Song Dataset. Secondarily, we will attempt to identify which attributes of a given song are the best predictors of a song's popularity.

We will assume the roles of consultants in the music industry whose clients include websites, radio stations and record labels. Our clients' revenue is driven by web clicks on songs that will result in advertising revenue for websites, holding the attention of radio-station listeners so that listeners can be exposed to more commercials, and song purchases at various portals that channel revenue to song producers. The common challenge in all of these transactions is determining what type of songs to recommend to a music consumer based on song attributes and the user's interests. Our job as consultants is to develop the analytics model a music recommendation engine will use to deliver song recommendations to commercial organizations and their customers.

We will experiment with select data mining methods and expect that the association-rule mining and clustering methods will be particularly useful for building the recommendation engine. We will also experiment with collaborative filtering data mining techniques.

The project is significant for two reasons: the recommendation engine will not only have commercial impact in the music industry but it may be possible to generalize our findings to other industries in terms of identifying a more effective methodology for implementing a recommendation engine.

Data Selection

Our team selected two related datasets for the data mining project: the Million Song Dataset and the Taste Profile Subset. Both datasets are curated by the Laboratory for the Recognition and Organization of Speed and Audio (LabROSA) at Columbia University. The datasets are available for download online at: http://labrosa.ee.columbia.edu/millionsong/.

The Million Song Dataset represents a free collection of song and audio metadata derived from one million contemporary popular music tracks. The songs span a wide variety of musical genres – from rock to blues to classical. The dataset does not include any actual audio data. It contains the audio metadata associated with songs such as the number of song measures, tempo, and loudness as well as key song attributes such as the song title, artist name, and release date.

The Taste Profile Subset contains song selections and play counts from real music listeners. The listener information was anonymized to protect privacy and the songs are all contained within the Million Song Dataset.

The Million Song Dataset was created by a partnership between LabROSA and a company called The Echo Nest. This company is focused on curating millions of songs for the purpose of delivering useful analytics to other businesses. For example, one of the company's core products is a song recommendation engine used by Internet music streaming companies like Spotify.

LabROSA and The Echo Next have made the Million Song Dataset available to researchers to encourage further development of large-scale analytical algorithms. These organizations hope that interested parties will push the boundaries of current Music Information Retrieval (MIR) technologies.

Data acquisition

The Million Song Dataset is comprised of one million individual data files consuming over 300GB of disk storage organized in a multi-directory hierarchy. Each data file contains the metadata associated with one song track – the artist name, release name, audio analysis data, etc. The dataset developers could not put a million files in a single folder, so they devised a file directory hierarchy to store the individual files. Each song track has an associated Echo Nest track ID, and this ID was used as the basis for the directory naming structure. Every song file is located within a directory folder named after the 3rd, 4th, and 5th letters of the corresponding track ID.

The data files in the Million Song Dataset are stored in a format that is not easily consumed by HANA – a data format called HDF5. HDF5 is a data storage format that was developed by NASA to support "large, heterogeneous, hierarchical datasets" (hdfgroup.org). Fortunately, the Million Song Dataset includes a SQLite version of song metadata containing a subset of the data fields included in the full dataset (see Appendix A for full field listing).

The SQLite song metadata table includes the following data fields:

Data field	Description
track_id	The Echo Nest track ID
title	Song title
song_id	The Echo Nest song ID
release	Music album the track is located on
artist_id	The Echo Next artist ID
artist_name	Name of the music artist

artist_familiarity	The Echo Nest artist familiarity value
artist_hotttness	The Echo Nest artist hotness value
duration	Length of the song in seconds
year	year the song was released

The Taste Profile Subset contains song selection and play counts for over one million users of The Echo Nest music service. The dataset includes over 48 million individual records. Each record, called a triplet, consists of three fields: user_id, song_id, play count. For example a set of records for a specific user may look like:

User ID	Song ID	Play count
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAKIMP12A8C130995	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAPDEY12A81C210A9	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBBMDR12A8C13253B	2
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBFNSP12AF72A0E22	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBFOVM12A58A7D494	1

It is possible to join the Million Song Dataset and the Taste Profile Subset using the song_id field. Each song in the Million Song Dataset may be related to 0 or many records in the Taste Profile Subset (see Fig. 1).

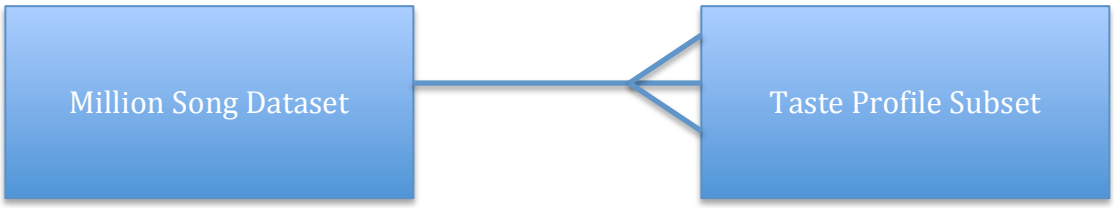

Fig 1. Basic E-R Diagram for analytical data

Data Pre-Processing

LabROSA invested significant time and effort organizing the Million Song Dataset and the Taste Profile Subset. Their efforts reduce the amount of time we have to spend fixing data records or dealing with missing data. Ultimately, we need to join the disparate datasets together for our data analysis.

The Million Song Dataset is stored in H5 files and SQLite database formats. The H5 files contain all of the song metadata and the SQLite database files contain a subset of the song metadata. We plan to use the SQLite databases for this project since these databases contain most of the metadata we need for analysis.

We will convert the SQLite database into a TSV text file. Once we have a text file, we will import this data into a HANA database table using the import tool in Eclipse. The Taste Profile Subset data is already stored in a TSV text file. We plan to import this TSV file into a HANA database and join the data with the Million Song Dataset using a view. We will use this view for the bulk of our analysis operations.

Existing Methods

There are two basic architectures for a recommendation system:

Content-Based Systems: These systems construct a recommendation based on new listeners' behavior. In a content-based system, the system stores a collection of all the relevant attributes based on past behaviors of a new listener. The system then compares the attributes of the new listener's profile against the attributes of a content object (i.e., song). Once the filtering process is complete, the algorithm determines whether or not the two items are similar. If they are similar, the content object is recommended to the listener.

Collaborative Filtering Systems: These systems focus on building a model that is based on the behaviors of previous listeners. The system can either group listeners who have the same music preferences together or it can create a model based solely on a new listener. When other listeners' behaviors are taken into consideration, the next song recommendation is established by using the information about the group to determine most likely trends. Through filtering through the listeners and narrowing the set down to only the listeners with similar habits, the system is able to predict a similar song or a new song the listener will most likely enjoy listening to. Collaborative Filtering Systems use methods such as Jaccard Similarity Coefficient and Cosine Similarity to build predictions.

Project Challenges

The central challenge in this project is producing a viable model for a recommendation engine. We plan to use two different methodologies: a content-based approach using market-basket analysis and a collaborative filtering approach. We understand how to create Association Rules using SAP HANA and we are well versed in content-based methodologies. We have no experience using a collaborative filtering approach and these types of algorithms are not covered in our class. We have not yet identified a tool to help us implement collaborative filtering analysis. We look at this challenge as a way to apply the knowledge we have gained from our course in a new application area, expanding our experience in the art of data mining.

References

Recommendation Systems (Chapter 9). Retrieved on March 24, 2015 from http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

Data Mining Author by Eko Indarto. Retrieved on March 24, 2015 from http://recommendersystems.readthedocs.org/en/latest/datamining.html

Microsoft Associaton Algorithm for SQL 2014. (2014) Retrieved on March 24, 2015 from https://msdn.microsoft.com/en-us/library/ms174916.aspx

Pearson Product-Moment Correlation. Retrieved on March 24, 2015 from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Efficient Top-N Recommendation for Very Large Scale Binary Rated Datasets. Retrieved on March 24, 2015 from http://www.math.unipd.it/~aiolli/PAPERS/MSD_final.pdf

Appendix A: Million Song Dataset Fields

Field name	Value	Description
artist_mbid	**db92a151…**	musicbrainz.org ID...
artist_mbtags	**shape = (4,)**	this artist received 4 tags on musicbrainz.org
artist_mbtags_count	**shape = (4,)**	raw tag count of the 4 tags this artist received on musicbrainz.org
artist_name	**Rick Astley**	artist name
artist_playmeid	**1338**	the ID of that artist on the service playme.com
artist_terms	**shape = (12,)**	this artist has 12 terms (tags) from The Echo Nest
artist_terms_freq	**shape = (12,)**	frequency of the 12 terms from TEN (number between 0 and 1)
artist_terms_weight	**shape = (12,)**	weight of the 12 terms from TEN (number between 0 and 1)
audio_md5	**bf53f8113…**	hash code of the audio used for the analysis by The Echo Nest
bars_confidence:	**shape = (99,)**	confidence value (between 0 and 1) associated with each bar
bars_start	**shape = (99,)**	start time of each bar according to TEN this song has 99 bars
beats_confidence	**shape = (397,)**	confidence value (between 0 and 1) associated with each beat
beats_start	**shape = (397,)**	start time of each beat according to TEN, this song has 397 beats
danceability	**0.0**	danceability measure of song (between 0 and 1, 0=>not known)
duration	**211.69587**	duration of the track in seconds
end_of_fade_in	**0.139**	time of the end of the fade in, at the beginning of the song,
energy	**0.0**	energy measure (between 0 and 1, 0 => not analyzed)
key	**1**	estimation of the key the song is in by The Echo Nest
key_confidence	**0.324**	confidence of the key estimation
loudness	**-7.75**	general loudness of the track
mode	**1**	estimation of the mode the song is in by The Echo Nest
mode_confidence	**0.434**	confidence of the mode estimation
release	**Big Tunes…**	album name from which the track was taken
release_7digitalid	**786795**	the ID of the release (album) on the service 7digital.com
sections_confidence	**shape = (10,)**	confidence value (between 0 and 1) associated with each section
sections_start	**shape = (10,)**	start time of each section according to TEN
segments_confidence	**shape = (935,)**	confidence value (between 0 and 1) associated with segment
segments_loudness_max	**shape = (935,)**	max loudness during each segment
segments_loudness_max_time	**shape = (935,)**	time of the max loudness during each segment
segments_loudness_start	**shape = (935,)**	loudness at the beginning of each segment
segments_pitches	**shape = (935, 12)**	chroma features for each segment (normalized so max is 1.)
segments_start	**shape = (935,)**	start time of each segment (~ musical event, or onset)
segments_timbre	**shape = (935, 12)**	MFCC-like features for each segment
similar_artists	**shape = (100,)**	a list of 100 artists (their Echo Nest ID) similar to Rick Astley
song_hotttnesss	**0.8642488**	this song had a 'hotttnesss' of 0.8 (on a scale of 0 and 1)
song_id	**SOCWJDB…**	The Echo Nest song ID
start_of_fade_out	**198.536**	start time of the fade out, in seconds, at the end of the song,
tatums_confidence	**shape = (794,)**	confidence value (between 0 and 1) associated with each tatum
tatums_start	**shape = (794,)**	start time of each tatum, this song has 794 tatums
tempo	**113.359**	tempo in BPM according to The Echo Nest
time_signature	**4**	usual number of beats per bar
time_signature_confidence	**0.634**	confidence of the time signature estimation
title	**Never Gonna..**	song title
track_7digitalid	**8707738**	the ID of this song on the service 7digital.com
track_id	**TRAXLZU1…**	The Echo Nest ID of this particular track
year	**1987**	year when this song was released, according to musicbrainz.org